Journal Pre-proof

Recurrent-DC: A deep representation clustering model for university profiling based on academic graph

Xiangjie Kong, Jiaxing Li, Luna Wang, Guojiang Shen, Yiming Sun, Ivan Lee

PII: S0167-739X(20)32996-4
DOI: https://doi.org/10.1016/j.future.2020.10.019
Reference: FUTURE 5887

To appear in: *Future Generation Computer Systems*

Received date : 13 June 2020
Revised date : 26 September 2020
Accepted date : 20 October 2020

Please cite this article as: X. Kong, J. Li, L. Wang et al., Recurrent-DC: A deep representation clustering model for university profiling based on academic graph, *Future Generation Computer Systems* (2020), doi: https://doi.org/10.1016/j.future.2020.10.019.

Recurrent-DC: A Deep Representation Clustering Model for University Profiling Based on Academic Graph

Xiangjie Kong[a,b], Jiaxing Li[b], Luna Wang[c], Guojiang Shen[a,*], Yiming Sun[b], Ivan Lee[d]

[a]College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China
[b]School of Software, Dalian University of Technology, Dalian 116620, China
[c]Institute of Science and Technology, Dalian University of Technology, Dalian 116024, China
[d]School of Information Technology and Mathematical Sciences, University of South Australia, Adelaide, SA 5001, Australia

Abstract

Universities play an important role in exploring new concepts and knowledge transfer. University research naturally forms heterogeneous graphs through all real-life academic communication activities. In recent years, there have been many large scholarly graph datasets containing web-scale nodes and edges. However, so far, for these graph data, characterizing research about university output is focusing on counting the volume or evaluating the excellence of research articles and providing a ranking. This paper proposes a novel University Profiling Framework (UPF) from the production and complexity point of view which is different from other straightforward solutions. The framework includes a novel Recurrent Deep Clustering Model (Recurrent-DC) for the learning of deep representations and clusters. In our model, successive operations in a clustering algorithm are expressed as steps in a recurrent process, stacked on top of representations output by a Stacked Autoencoder (SAE). Our key idea behind this model is that good representations for university clustering task-specific problem can be learned over multiple timesteps. Experimental results illustrate the stability and effectiveness of the proposed model comparing with the other deep clustering and classical clustering methods.

Keywords: academic graph, university profiling, deep representation learning, deep clustering, Stacked Autoencoder

1. Introduction

Universities play an important role in exploring new concepts and innovation by research, in addition to knowledge transfer through higher education. University academic research communications can be naturally modeled as heterogeneous graphs. Heterogeneous graphs have been commonly used for abstracting and modeling complex systems, in which objects of different types interact with each other in various ways. Academic Graph has received a lot of attention in recent years as an important example of heterogeneous graphs. For example, Microsoft Academic Graph (MAG) [1], contains six types of entities: field of study, author, institution (affiliation of author), paper, venue (journal and conference series, e.g. WWW, SIGIR, KDD, etc.) and event (conference instances). Different types of relationships between entities are also included. These entity relationships are rather intuitive. For instance, the fact that papers get published in journals/conferences justifies the edge between paper and venue nodes in the graph.

Various scientific algorithms have been developed to quantify and assess academic institutes and provide rankings [2, 3, 4], using proprietary or public accessible publication records of the academic graph. However, ranking universities is a challenging task because each institution has its own particular mission. Each institution has its focus and offers different academic programs. Institutions can also differ in size and have varying amounts of resources at their disposal. In addition, each country has its own history and higher education system which can impact the structure of their colleges and universities and how they compare to others. Universities usually consist of complex research disciplines in different faculties or divisions and conduct different research in different fields. The analysis of the whole university is a challenging task. The main contribution of this article is the use of complexity-based institutional evaluation indicators and the development of corresponding deep clustering algorithms. The traditional method is to count the output or count the number of outstanding works. For example, count the number of papers published by an institution in a year, and count the number of papers published by an institution in top journals in a year. In this paper, evaluating excellence is no longer a simple matter of counting outstanding works. We used a complexity-based indicator to replace the number of outstanding works. The complexity of papers published by institutions is not only related to the excellent degree of the papers, but also to the proportion of papers published. Instead of ranking universities, more and more investigators choose to calculate various scientific indicators first and then cluster universities [5]. There are much research about different scientific indicators [6, 7] but few investigate about university profiling task-specific clustering algorithm.

Clustering is one of the most fundamental tasks in data mining and machine learning, with an endless list of applications.

*Corresponding author
Email address: gjshen1975@zjut.edu.cn (Guojiang Shen)

It is also a notoriously hard task, whose outcome is affected by a number of factors, including data acquisition and representation, preprocessing, clustering criterion, etc. Since its introduction in 1957 by Lloyd (published much later in 1982) [8], K-means has been extensively used either alone or together with suitable preprocessing, due to its simplicity and effectiveness. K-means is suitable for clustering data samples that are evenly spread around some centroids [9]. Many real-life datasets do not exhibit this specific structure. And many scientific indicators represent special data features that usually do not exhibit this specific structure. This task-specific issue limits the classic clustering algorithm performance.

In recent years, motivated by the success of deep neural networks (DNNs) in supervised learning, unsupervised deep learning approaches are now widely used for representation learning prior to clustering. For example, the Stacked Autoencoder (SAE) [10], make use of DNNs to learn nonlinear mappings from the data domain to low-dimensional latent spaces. These approaches [11] treat their deep neural networks as a preprocessing stage that is separately designed from the subsequent clustering stage. The hope is that the latent representations of the data learned by these deep neural networks will be naturally suitable for clustering. However, since no task-specific objective is explicitly incorporated in the learning process, the learned deep neural networks do not necessarily output data that are suitable for clustering. Besides, there are some approaches [9, 12] attempt their deep neural networks and clustering part optimizing jointly to get a better result. They hope fully use the power of stochastic gradient descent algorithm not only in optimizing deep neural network parameters but also in the clustering assignment. However, for university profiling task clustering, without explicit learning normalization, optimizing jointly will not necessarily output results that are suitable for university clustering task - as will be seen in our experiments.

In this paper, we propose a task dependence model, which alternates between two steps recurrently: updating the cluster assignment given the current representation parameters and updating the representation parameters given the current clustering result. To the best of our knowledge, no prior effort has been made to address the scientific features for university clustering by exploiting deep representation learning. Specifically, we cluster data representations using K-means clustering and represent observable data via activations of a SAE. We regard the university profiling problem as a clustering problem. The production and complexity of each university are modeled as joint vectors. We design an efficient algorithm to optimize the process of vector representation in the joint space and conduct clustering. We conduct various experiments to evaluate the effectiveness of our model. Results show that this method outperforms the other models.

In summary, the main contributions of this paper can be summarized as follows:

- We propose a new Recurrent Deep Clustering Model (Recurrent-DC). Since the deep representation learning and clustering are recurrent processes, it has higher clustering accuracy and better stability than other state-of-the-art models.
- We propose a new University Profiling Framework (UPF) for characterizing scientific research institutions in exploring academic graph dateset, which transforms the traditional depiction of excellence into the quantification of complexity indicators.
- We select the data of top universities from ARWU and apply our framework and model on the Microsoft Academic Graph. Then we find the positive relationship between university research production and university research complexity of different university groups.

The rest of the paper is organized in the following way. Section 2 lays out the theoretical dimensions of the research. Section 3 formally formulates the problem and presents the overall architecture of the proposed solution. Section 4 describes the experimental setups and presents results to illustrate the effectiveness of Recurrent-DC. Besides, we also analyze the results and present the findings focusing on the application and visualization of the research in section 5. Finally, Section 6 concludes our work and discusses the limitation and the promising future directions.

2. Related work

The related work has been divided into two parts. The first part deals with the institutional academic assessment in exploring academic graph. The second part presents focuses on the process from clustering to deep representation clustering.

2.1. Institutional assessment in exploring academic graph

With the dramatic growth of the research output volume and the increase of competition between universities, academic graph analysis has been an active research area [13, 14, 15, 16, 17, 18, 19, 20]. Comparing to others, research publications are the most popular benchmark to assess the performance of a university [21, 22], subsequently, influence the university education fund and research grants even the attractiveness to top talents to join the university. Graph analysis as an important technical has appeared a lot of work [23, 24, 25, 26, 27]. Thus, quantifying research publications to reflect the true influence is crucial for university research output analysis.

Popular university ranking systems, such as Shanghai Jiao Tong Academic Ranking of World Universities (ARWU), Quacquarelli Symonds World University Ranking (QS), Times Higher Education World University Rankings (THE), and Best Global Universities Rankings from US News, consider research outcome as a crucial factor in ranking universities around the world. The present research outcome quantifying methods are the most widely used methods for research assessment.

As universities usually consist of complex research disciplines in different faculties or divisions and conduct different research in different fields, the analysis of research complexity has attracted attention in the research community. Li et al. [6, 7] adopted the Research Complexity Index for analyzing the

research complexity of universities based on research publications. They use research publications to measure the diversity and ubiquity of universities' research output.

Tracing the development of institutional academic assessment, there are many research about different scientific indicators but little investigate about university profiling task-specific clustering algorithm. The framework proposed in this article is an in-depth profile of university research institutions. The framework of this article is meaningful in many aspects, such as the formulation of university discipline development plans. The subject development of the university can refer to the research results of this article. This situation is similar to business in the real world, in which knowledge is treated as a commodity, and university leaders usually define strategic research directions to maintain or expand their competitiveness [6, 7]. High achievements in research activities help universities to obtain educational and research funding, attract industrial cooperation or applied research services, and recruit top talents to join research projects. Therefore, for most research institutions, it is crucial to determine strategic research priorities to help improve reputation.

2.2. From traditional clustering to deep clustering

Given a set of data samples $\{\boldsymbol{x}_i\}_{i=1,\ldots,N}$ where $\boldsymbol{x}_i \in \mathbb{R}^M$, the task of clustering is to group the N data samples into K categories. Clustering algorithms can be broadly categorized into partitional and hierarchical approaches. The most well-known method in partitional clustering approaches is K-means [28], which minimizes the sum of square errors between data points and their nearest cluster centers. K-means approaches the clustering task by optimizing the following cost function:

$$\min_{\boldsymbol{M}\in\mathbb{R}^{M\times K},\{\boldsymbol{s}_i\in\mathbb{R}^K\}} \sum_{i=1}^{N} \|\boldsymbol{x}_i - \boldsymbol{M}\boldsymbol{s}_i\|_2^2, \tag{1}$$

where $\boldsymbol{s}_i$ is the assignment vector of data point i which has only one non-zero element, $s_{j,i}$ denotes the jth element of s_i, and the kth column of $\boldsymbol{M}$, that is to say, $\boldsymbol{m}_k$, denotes the centroid of the kth cluster. Related ideas form the basis of a number of methods, such as Affinity Propagation (AP) [29], Spectral Clustering (SC) [30], and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [31]. As for hierarchical clustering methods, CURE [32], BIRCH [33], etc. are typically based on hierarchy.

Many works use raw data intensity or hand-crafted features combined with conventional clustering methods. Such as K-means works well when the data samples are evenly scattered around their centroids in the feature space. However, some observable data especially high-dimensional data are in general not clustering-friendly in the feature space for some specific tasks. Recently, representations learning using deep neural networks have presented significant improvements over hand-designed features on many tasks. However, these approaches rely on supervised learning with large amounts of labeled data to learn rich representations. A number of works have focused on learning representations from unlabeled image data. One class of approaches cater to reconstruction tasks, such as Autoencoders [34], Deep Belief Networks (DBN) [35], etc.

In practice, a number of works have explored combining clustering tasks with representation learning. Using representations learning as a pre-processing to reduce the dimension or get better representations of observable data $\boldsymbol{x}_i$ to a much lower dimensional space or task-specific space and then apply K-means usually gives better results. In addition to the classic dimensionality reduction methods such as PCA or NMF that essentially learn a linear generative model from the latent space to the data domain, nonlinear representation learning approaches such as Autoencoders [34] and Deep Belief Networks [35] are also widely used as pre-processing before K-means or other clustering algorithms. For example, the authors [36] proposed to learn a non-linear embedding of the undirected affinity graph using Stacked Autoencoder, and then ran K-means in the embedding space to obtain clusters.

Instead of using representation learning as a pre-processing, joint representation leaning and clustering was also considered in the literature [37, 9, 12]. They seek powerful non-linear transforms, such as Stacked Autoencoder and Convolutional Neural Network (CNN), to model the observable data generating process, while at the same time make use of the joint representation learning and clustering idea. The idea of these literature [37, 12] is to connect a clustering module to the output layer of a deep neural network, and jointly learn neural network parameters and clusters. This idea seems reasonable but is problematic. The global optimal solution can always be achieved if the output of their deep neural network is zero. Other trivial solutions are simply mapping arbitrary data samples to tight clusters, which will lead to a small value of loss. But this could be far from being desired since there is no provision for respecting the data samples. To fix those problems, Deep clustering Network [9] making latent features are also responsible for reconstructing the input. However, jointly optimizing clustering loss and reconstructing loss may result in the fluctuation of data representations. Thus, they are not suitable for all types of clustering problems.

3. University Profiling Framework

In this work, we focus on characterizing universities in exploring academic graph. Different from the existing study, we utilize new scientific indicators in our University Profiling Framework to characterizing university rather than traditional excellence or volume indicators. The new scientific indicators contain Research Production Index (RPI), Productivity Value (PV), Research Complexity Index (RCI) and Opportunity Value (OV). And the new Recurrent Deep Clustering Model was developed to get better performance. Fig. 1 is an overview of our University Profiling Framework. To clearly illustrate the framework, we first describe the scientific indicators used in our UPF framework. And then we will present the details of how to construct the Recurrent-DC and the optimization progress.

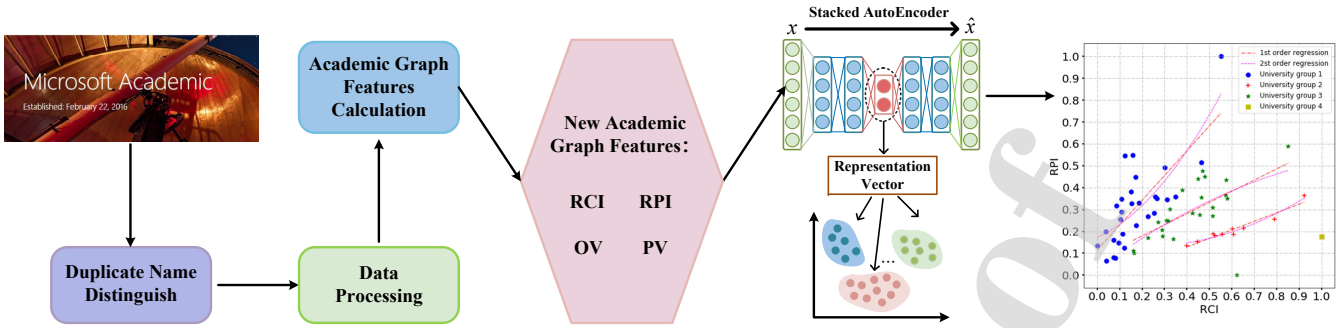


Figure 1: The structure of University Profiling Framework

3.1. Framework formulation

3.1.1. Research production and productivity

The annual number of research publications produced in a university is a standard production indicator. While research publication is a popular means of quantifying research outcome, publication record of different research disciplines is used in this study. We utilize Microsoft Academic Graph [1, 38, 39] to calculate university research production and productivity because of its availability as an open data set. This paper has chosen the 2018 collection for analysis. Top 100 universities listed on the ARWU ranking have been selected as the initial candidates for the study. For the reason that MAG considers all the brunch campuses, we remove the university name that is not specifically referred. Among all of the top 100 universities during the periods of 2003-2017, there are 64 universities always on the top 100 list of ARWU. Those 64 universities have been chosen as the research analysis target in this paper, assuming top universities have established research infrastructures and a dramatic change in research production in adjacent years is less likely to happen. And the detailed data set processing process can be seen in the next section.

Let u denotes a university, and $p(u)$ denotes a paper published in journals that has at least one author from university u. The measure of a university's research production $L(u)$ can be formulated as:

$$L(u) = \sum_{i \in p(u)} i \tag{2}$$

The Research Production Index of a university, $RPI(u)$, is calculated according to the average value of $L(u)$ per year during the period we selected.

Let a denotes an author, and $a(u)$ denotes an author who belongs to university u. The measure of a university's researcher $H(u)$ can be formulated as:

$$H(u) = \sum_{j \in a(u)} j \tag{3}$$

We define the $RNI(u)$ as a University's Number of Researcher Index which is calculated according to the average value of $H(u)$ per year during the period we selected. We identify and compare the research performance of the top-ranked universities around the world during the recent period 2015-2017.

However, it is important to remark that the number represents only estimations of the real quantities, because there maybe few researchers have not published papers during the period we selected.

To further explore the scientific production efficiency of a university, Research Productivity Value $PV(u)$ is defined as the ratio of university average annual number of research publications $RPI(u)$ and the university's number of researcher per year $RNI(u)$, that is:

$$PV(u) = \frac{RPI(u)}{RNI(u)}. \tag{4}$$

A higher Productivity Value indicates that the university researcher is more efficient in producing research publications. It should be noted that the Productivity Value focusing on the efficiency to publish the paper and it does not guarantee the quality of the published article.

3.1.2. Research complexity and opportunity

Besides considering a university's research production, it is also important to consider the research specialization of a university. In order to study the complexity and diversity of university research, Research Complexity Index and Opportunity Value [6] are adopted for analyzing university research specialization of different disciplines. In our study, we use the Microsoft Academic Graph [1, 38, 39] fields of study term which labels research paper with different topics. Although, there are a number of multi-disciplinary papers related to more than one research field. Let u denotes a university, and f denotes a research field. Let m denotes the number of all disciplines that a paper related to and $p(u, f)$ denotes a paper published. A paper may be labeled against one or more labels, related to single or multi-disciplinary research topics. Thus, $m \geq 1$ and $m \in Z$.

Besides, considering a paper may be published in different journals with different influences, a weighting factor $n_{p(u,f)}$ needs to be applied. This is the important step to increase the weight factor. A weighting factor can better show the influence of different articles. $n_{p(u,f)}$ denotes that for a paper $p(u, f)$ published in journals cited by n paper where $n \in Z$ and $n \geq 0$. Let publication matrix $\boldsymbol{P}_{uf}$ denote the summation over all papers

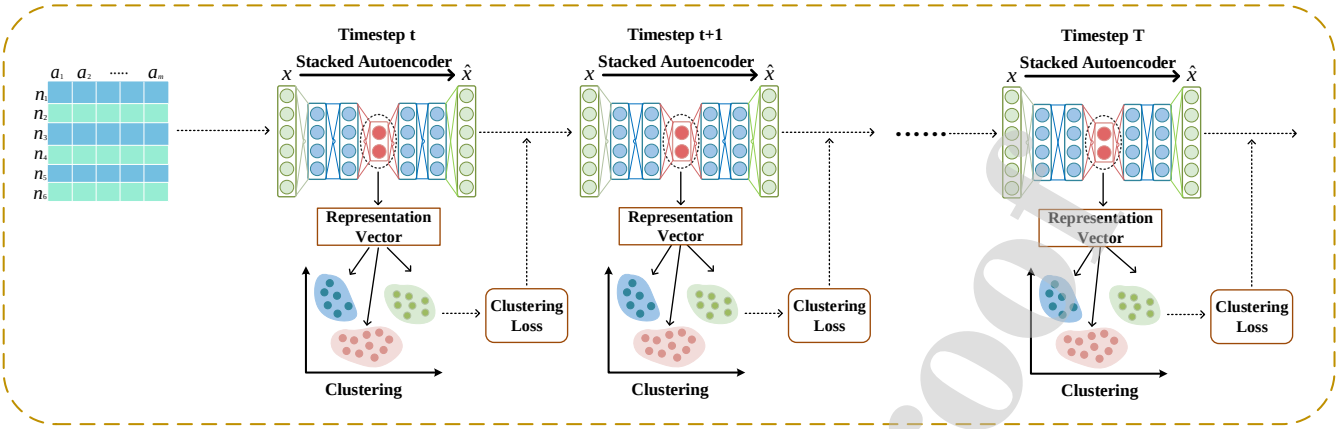


Figure 2: The structure of Recurrent Deep Clustering Model

$p(u, f)$ in research field f and has at least one author from university u, multiplied by a normalization factor $m_{p(u,f)}$ according to the number of research fields related to the paper.

$$\boldsymbol{P}_{uf} = \sum_{p(u,f)} \frac{n_{p(u,f)}}{m_{p(u,f)}}. \tag{5}$$

From $\boldsymbol{P}_{uf}$, the Revealed Symmetric Comparative Advantage (RSCA) can be calculated. The Revealed Symmetric Comparative Advantage is a symmetric version of Revealed Comparative Advantage (RCA) which is the percentage of academic publications of a particular discipline in a university, divided by the percentage of all papers by the university over all academic publications [6]. Thus, if the university has a strong focus on a particular research discipline, then the nominator term increases hence increase the overall RCA value.

The Research Complexity Index of a university, $RCI(u)$, is calculated from

$$RCI(u) = \frac{\boldsymbol{K}_u - \overline{\boldsymbol{K}_u}}{\sigma(\boldsymbol{K})}. \tag{6}$$

$\boldsymbol{K}_u$ denotes the eigenvector of iterative matrix associated with the second largest eigenvalue and $\sigma(\boldsymbol{K}_u)$ denotes the standard deviation of the $\boldsymbol{K}$ vector which can be calculated from RSCA according to [6].

Similarly, Field Complexity Index $FCI(f)$ is able to model the complexity of research fields can also be calculated from RSCA. In the following equations, f and f represent different research fields, and d represents capability distance, M indicating whether an academic institute possesses revealed symmetric comparative advantage for a research field.

The Opportunity Value of a university u [6] is defined as:

$$OV(u) = \sum_{f'} (1 - d(u, f'))(1 - M(u, f'))FCI(f'). \tag{7}$$

A higher Opportunity Value indicates that the institute has more research fields in close proximity or its research fields are more complex.

3.2. Recurrent Deep Clustering Model

We are motivated to model the relationship between the observable data x and its clustering task friendly latent representation h using a nonlinear mapping:

$$\boldsymbol{h}_i = \boldsymbol{f}(\boldsymbol{x}_i; \mathcal{W}), \tag{8}$$

where $f(\boldsymbol{x}_i; \mathcal{W}) : \mathbb{R}^M \to \mathbb{R}^R$ denotes the network function output given data sample $\boldsymbol{x}_i$, $\mathcal{W}$ collects the network parameters. In this work, we propose to employ a deep neural network as network function, since deep neural network has the ability of approximating [40] continuous mapping using a reasonable number of parameters.

In the realm of unsupervised deep neural network, there are several well-developed approaches for reconstruction e.g., the Stacked Autoencoder is a popular choice for serving this purpose. The power of deep learning lies in its ability to learn multiple expressions of raw data layer by layer. Each layer is based on the expression features of the previous layer, extracting more abstract and more suitable for complex features, and then doing some classification tasks. Stacked Autoencoder actually does such a thing, which is composed of multiple automatic encoders stacked in series. The purpose of the stacked multi-layer autoencoder is to extract the high-order features of the input data layer by layer. In this process, the dimensionality of the input data is reduced layer by layer, and a complex input data is transformed into a series of simple high-order features. Then input these high-level features into a classifier or clusterer for classification or clustering. To prevent trivial representations such as all-zero vectors, SAE uses a decoding network $g(\boldsymbol{h}_i; \mathcal{Z})$, where $\mathcal{Z}$ collects the decoding network parameters, to map the $\boldsymbol{h}_i$ back to the data domain and requires that $g(\boldsymbol{h}_i; \mathcal{Z})$ and $\boldsymbol{x}_i$ match each other well under some metric, e.g., mutual information or least squares-based measures. In a general way, SAE has the following cost function:

$$\min_{\mathcal{W},\mathcal{Z}} \sum_{i=1}^{N} \mathcal{L}(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i; \mathcal{W}); \mathcal{Z}), \boldsymbol{x}_i). \tag{9}$$

The function $\mathcal{L} : \mathbb{R}^M \rightarrow \mathbb{R}$ is a certain loss function that measures the reconstruction error. In this work, we adopt the least-squares loss $\mathcal{L}(x, y) = \|x - y\|_2^2$; other choices such as $\mathcal{L}_1$-norm based fitting and the KL divergence can also be considered.

We want to connect a clustering module to the deep neural network and jointly learn deep neural network parameters and clusters. Specifically, the approaches look into an optimization problem of the following form:

$$\min_{\mathcal{W},\Theta} \sum_{i=1}^{N} c(\boldsymbol{f}(\boldsymbol{x}_i; \mathcal{W})\Theta), \tag{10}$$

where Θ denotes parameters of some clustering model. In this work, we adopt the K-means clustering model and Θ stands for the centroids $\boldsymbol{M}$ and assignments $\{\boldsymbol{s}_i\}$. The c in Equation (10) denotes the clustering loss. In this work, we adopt clustering accuracy: ACC [41] as the clustering loss. This measuring metrics are commonly used in the clustering literature:

$$\max_{m} \frac{\sum_{i=1}^{N} \mathbf{1}\{l_i = m(\boldsymbol{s}_i)\}}{N}, \tag{11}$$

where l_i is the ground-truth label, $\boldsymbol{s}_i$ is the cluster assignment produced by the algorithm, and m ranges over all possible one-to-one mappings between clusters and labels. Intuitively this metric takes a cluster assignment from an unsupervised algorithm and a ground truth assignment and then finds the best matching between them. The best clustering accuracy makes the lowest loss of c in Equation (10).

Our key insight is that deep representation clustering can be interpreted as a recurrent process in the sense that it learns discriminative representations over multiple timesteps. Based on this insight, we propose a recurrent model to combine the clustering and representation learning processes.

Fig.2 presents the network structure corresponding to the formulation in Equation (10) . Comparing to other networks, our features representations are also responsible for reconstructing the input, having provision for respecting the data samples. At timestep t, data $\boldsymbol{x}$ are first fed into the SAE to get the representation vector. And then the representation vector gets into clustering procedure to calculate clustering loss. At the next timestep $t+1$, data $\boldsymbol{x}$ are fed into the SAE with the best network parameters depending on the clustering loss of timestep t.

In a typical Recurrent Neural Network (RNN), one would unroll all timesteps at each training iteration. In our case, that may get not reliable representations of SAE at the beginning. In this work, we split the overall T timesteps into multiple periods and unroll one period at a time. The intuitive reason we unroll one period an a time is that the representation of the SAE at the beginning is not suitable for our clustering. We need to update SAE parameters to obtain more discriminative representations for the following clustering processes. In each period, we update SAE parameters for a fixed number of iterations and update clusters at the end of the period. In our case, one timestep is one epoch of going through all the samples. Other choices for epochs in one timestep can also be considered.

In our experiments, at whole timestep T, $\lambda \geq 0$ is a regularization parameter which balances the overall timesteps versus finding appropriate latent representations. And the structure of the decoding networks is a mirrored version of the encoding network, and for both the encoding and decoding networks, we use the rectified linear unit (ReLU) [42] activation-based neurons.

3.3. Optimization

Optimization is highly non-trivial since the cost function and the clustering constraints are non-convex. In addition, there are scalability issues that need to be taken into account. In this section, we propose a pragmatic optimization procedure including:

3.3.1. Alternating stochastic optimization

It is very challenging to handle the non-convex problem of cost function and clustering constraints. The Alternating Stochastic Optimization commonly used stochastic gradient descent (SGD) algorithm cannot be directly applied to optimize recurrent progress because the block variable $\boldsymbol{s}_i$ is constrained on a discrete set. Our idea is that optimization can be solved in an alternating recurrent process. Specifically, we propose to optimize the subproblems with respect to one of $(\boldsymbol{M}, \{\boldsymbol{s}_i\})$ and $(\mathcal{W}, \mathcal{Z})$ while keeping the other two sets of variables fixed.

3.3.2. Update neural network parameters

For fixed clustering parameters $(\boldsymbol{M}, \{\boldsymbol{s}_i\})$, the subproblem with respect to $(\mathcal{W}, \mathcal{Z})$ is to train an SAE. We can take advantage of the mature tools for training deep neural network, e.g., back-propagation based SGD and its variants. To implement SGD for updating the network parameters, we look at the problem with respect to the incoming data $\boldsymbol{x}_i$:

$$\min_{\mathcal{W},\mathcal{Z}} L^i = \mathcal{L}(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i; \mathcal{W}); \mathcal{Z}), \boldsymbol{x}_i). \tag{12}$$

The gradient of the above function over the network parameters is easily computable, i.e., $\triangledown_{\mathcal{X}} L^i = \frac{\partial \mathcal{L}(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i;\mathcal{W});\mathcal{Z}),\boldsymbol{x}_i)}{\partial \mathcal{X}}$ where $\mathcal{X} = (\mathcal{W}, \mathcal{Z})$ is a collection of the network parameters and the gradients $\frac{\partial \mathcal{L}}{\partial \mathcal{X}}$ and $\frac{\partial \boldsymbol{f}(\boldsymbol{x}_i)}{\partial \mathcal{X}}$ can be calculated by back-propagation. Strictly speaking, what we calculate here is the subgradient with respect to $\mathcal{X}$ since the ReLU function is nondifferentiable at zero. Then, the network parameters are updated by

$$\mathcal{X} \leftarrow \mathcal{X} - \alpha \triangledown_{\mathcal{X}} L_i, \tag{13}$$

where $\alpha > 0$ is a diminishing learning rate.

3.3.3. Update clustering parameters

For the fixed neural network parameters, we have latent representation $\boldsymbol{h}_i$ of observable data $\boldsymbol{x}_i$ as shown in Equation (8).

The assignment vector of data representations $\boldsymbol{s}_i$ and the centroid of the cluster $\boldsymbol{M}$ can be naturally updated by optimizing the minimum distance as follows:

$$s_{j,i} = \begin{cases} 1, & \text{if } j = \arg\min_{k=\{1,\ldots,K\}} \|\boldsymbol{h_i} - \boldsymbol{m}_k\|_2, \\ 0, & \text{otherwise,} \end{cases} \tag{14}$$

where $\boldsymbol{s}_i$ is the assignment vector of data point i which has only one non-zero element, $s_{j,i}$ denotes the jth element of $\boldsymbol{s}_i$, and the

Algorithm 1 Recurrent deep clustering model algorithm

Input:

$\boldsymbol{x}$: = collection of data;

l^*: = target labels of clusters;

Output:

$\boldsymbol{s}^*, \mathcal{X}^*$: = final labels and SAE parameters;

Initialization: randomly initiate cluster labels and SAE parameters $\boldsymbol{s}, \mathcal{X}$ {Perform T timesteps over the data}

for $t = 1 : T$ **do**

  Calculate reconstruction loss L based on $\mathcal{L}(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i; \mathcal{W}); \mathcal{Z}), \boldsymbol{x}_i)$

  Calculate the gradient over network parameters $\mathcal{X}$ based on $\nabla_{\mathcal{X}} L^i = \frac{\partial \mathcal{L}(\boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}_i; \mathcal{W}); \mathcal{Z}), \boldsymbol{x}_i)}{\partial \mathcal{X}}$

  Update network parameters $\mathcal{X}$ based on $\mathcal{X} \leftarrow \mathcal{X} - \alpha \nabla_{\mathcal{X}} L_i$

  For fixed parameters $\mathcal{X}$, calculate the latent representation $\boldsymbol{h}$

  For fixed parameters $\mathcal{X}$ and latent representation $\boldsymbol{h}$, calculate the cluster assignment vector and centroid of the cluster $\boldsymbol{s}$, $\boldsymbol{M}$

  Update clustering assignment $\boldsymbol{s}$ based on $\arg\min_{k=\{1,\ldots,K\}} \|\boldsymbol{h_i} - \boldsymbol{m}_k\|_2$

  Update clustering centroids $\boldsymbol{M}$ by Equation (15)

end for

kth column of $\boldsymbol{M}$, namely $\boldsymbol{m}_k$, denotes the centroid of the kth cluster. The clustering can be summarized by optimizing the following cost function:

$$\min_{\boldsymbol{M} \in \mathbb{R}^{M \times K}, \{\boldsymbol{s}_i \in \mathbb{R}^K\}} \sum_{i=1}^{N} \|\boldsymbol{h}_i - \boldsymbol{M}\boldsymbol{s}_i\|_2^2. \tag{15}$$

The overall architecture of the algorithm is summarized in Algorithm 1. The detailed steps of calculating reconstruction loss are described in the previous Stacked Autoencoder description. Calculating the gradient over network parameters is a loss function formula based on neural networks. Updating network parameters are based on the previous standard iteration formula. Note that a timestep corresponds to the pass of all data samples concurrently through the network.

4. Experiments

In this section, we will present the details about how we carry out the experiments to certify the accuracy and the effectiveness of our model. We describe how to get the suitable experimental dataset from Microsoft Academic Graph. We compare the proposed Recurrent-DC model with various clustering methods including K-means, Affinity Propagation, Spectral Clustering, Density-Based Spatial Clustering of Applications with Noise, Stacked Autoencoder followed by K-means (SAE+KM) and Deep clustering Network.

4.1. Experimental setup

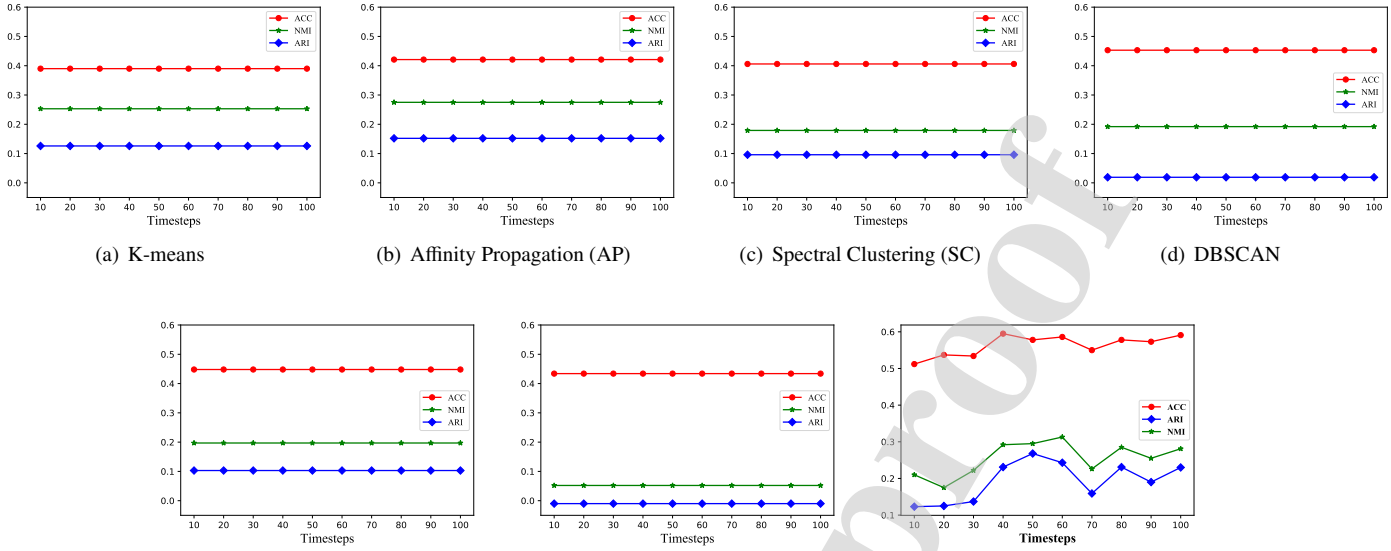
4.1.1. Datasets

We use the Microsoft Academic Graph [1, 38, 39], which is a widely used dataset containing scientific publication records to analyze research output produced by a university because of its availability as an open data set. Comparing with other popular open data collections such as DataBase systems and Logic Programming (DBLP) for computer science and American Physical Society (APS) for physics, Microsoft Academic Graph covers a wide range of research disciplines so it can be used to generally characterize universities. In comparison with other multidisciplinary datasets such as ISI Web of Science (WoS) and Google Scholar, the full collection of the MAG dataset is open to the public to access and download. The MAG dataset was the official dataset for the 2016 KDD CUP competition. Plenty of articles [43, 44, 45] suggest the MAG dataset becomes the data source of choice with a more structured approach to data presentation. This paper has chosen the 2018 collection for analysis.

The format of academic publication may be considered semi-structured data, and different publishers specifying different paper formats which make it challenging to process these data. It is worth noting that the concept of institutional complexity we adopt is not complicated in the traditional sense. For example, the complexity of an organization is not directly proportional to the disciplines of the organization. The more disciplines an organization has and the stronger its competitiveness, the higher the complexity of the organization. This is also the significance of replacing the traditional evaluation excellent level mentioned in the article. MAG automatically parse publication details such as authors and universities through specific algorithms [1]. In this study, university information is required and it is observed that some punctuation, such as the hyphen in university name, is removed in the MAG dataset. For instance, the hyphen in "University of Wisconsin-Madison" is removed in the MAG dataset. In addition, well-perceived abbreviations such as "ETH Zurich" for "Swiss Federal Institute of Technology Zurich" are considered. Thus, an adjustment needs to be made for the university name. To help clean up the dataset, a manual inspection of the selected affiliation names used in this paper is conducted.

4.1.2. Data processing

Top 100 universities listed on the ARWU ranking have been selected as the initial candidates for the study. It should be noted that the university name in MAG may be different from the name in the Shanghai Jiao Tong Academic Ranking of World Universities. For example, the 24th in ARWU 2017 is the University of Michigan-Ann Arbor. However, the university name in the MAG is 'University of Michigan'. MAG considers the university of Michigan system but not specific to the flagship campus. So in the MAG, articles published by other campuses are also included in the name 'University of Michigan'. For the reason that MAG considers all the brunch campuses, we remove the university name that are not specifically referred. Among all of the top 100 universities during the periods of 2003-2017, there are 64 universities always on the top 100 list of ARWU.



(a) K-means (b) Affinity Propagation (AP) (c) Spectral Clustering (SC) (d) DBSCAN

(e) SAE+KM (f) DCN (g) Recurrent-DC

Figure 3: Clustering performance of different methods over timesteps.

Those 64 universities have been chosen as the research analysis target in this paper, assuming top universities have established research infrastructures and a dramatic change in research production in adjacent years is less likely to happen. It should be noted that the choice of the top 100 universities of ARWU is not a necessity. The framework proposed in this paper is able to characterize any university group identified by various standards or other classification methods like top universities in different countries.

Before applying Recurrent Deep Clustering model on the MAG dataset to gain the academic profiling of universities, the dataset attributes pose a challenge. The challenge is that the data set does not classify universities as ground truth for the new scientific indicators we adopt. In order to solve this challenge and explore the relationship between Research Complexity Index and Research Production Index of university, we regard RCI / RPI as a new feature as one of the data features for clustering. Then use k-means algorithm for clustering. The cluster number is the optimal group determined by commonly used clustering indicators. At the same time, the clustering comparison is divided into 3, 4, 5, 6 groups. According to the cluster evaluation index (Silhouette Coefficient, Davies Bouldin Score), we selected the results of 4 clusters as ground truth.

4.1.3. Baselines

We compare the proposed Recurrent-DC with a variety of baseline methods:

- **K-means**: The classic K-means algorithm [8].
- **Affinity Propagation (AP)**: The classic AP algorithm [29].
- **Spectral Clustering (SC)**: The classic SC algorithm [30].
- **Density-Based Spatial Clustering of Applications with Noise (DBSCAN)**: It is a density-based clustering algorithm. Given a set of points in some space, it groups together points that are closely packed together [31].
- **Stacked Autoencoder followed by K-means (SAE+KM)**: This is a two-stage approach. We use SAE for representation learning first and then apply K-means.
- **Deep Clustering Network (DCN)**: DCN performs joint DNN and clustering, where the loss function contains not only clustering loss but also penalty on reconstruction [9].

For the experiment, we select the baselines that are considered most competitive and suitable for the application.

4.1.4. Evaluation metrics

Evaluating the performance of a clustering algorithm is not as trivial as counting the number of errors or the precision and recall of a supervised classification algorithm. In particular, any evaluation metric should not take the absolute values of the cluster labels into account.

We use standard metrics to evaluate clustering performance. Specifically, we use the following three metrics: clustering accuracy (ACC), adjusted Rand index (ARI) and normalized mutual information (NMI). In short, the above three metrics are commonly used in clustering literature, and each has advantages and disadvantages. However, using them together is sufficient to prove the effectiveness of the clustering algorithm. Note that NMI and ACC are in the range of zero to one, one of them is the perfect clustering result, and zero is the worst clustering result. ARI is a value between -1 and 1, one of them is the best clustering performance, and the minus one is the worst result.

4.2. Results and analysis

To evaluate the effectiveness of the proposed method Recurrent-DC on data representations clustering, we conduct a series of experiments from the following aspects. First, in order to illustrate how much Recurrent-DC can improve the embedding representation and clustering, we compare our proposed model with state-of-the-art deep clustering models and classic clustering models mentioned above.

Table 1: Evaluation of different methods

Methods	ACC	ARI	NMI
kmeans	0.390	0.126	0.253
AP	0.421	0.152	0.275
SC	0.406	0.096	0.179
DBSCAN	0.453	0.019	0.192
SAE+KM	0.448	0.103	0.197
DCN	0.434	-0.010	0.052
Recurrent-DC	**0.578**	**0.268**	**0.295**

4.2.1. Effectiveness evaluation

TABLE 1 presents the results of comparing Recurrent-DC with all baselines. Note that, parameters setting in the methods of DCN are decided by the maximum value of accuracy. From the results, we can see that the task of university profiling clustering achieves consistent and significant improvements by taking advantage of our proposed model. It demonstrates the effectiveness of Recurrent-DC in the task of university profiling clustering in exploring big scholar data.

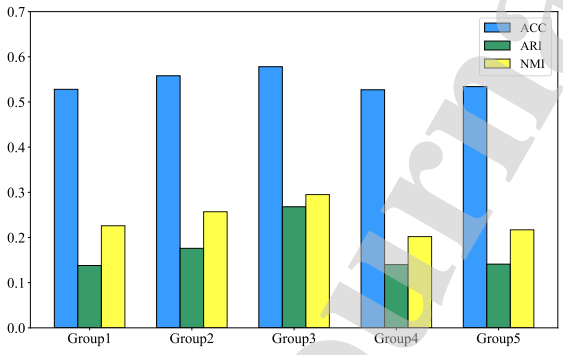


Figure 4: Recurrent-DC performance with different hidden layers in the Stacked Autoencoder.

Besides, we have drawn the trend of evaluation metrics of different methods over timesteps from 10 to 100. From Fig. 3, one can see a clear overall ascending trend of every evaluation metric in our proposed method Recurrent-DC. In Fig. 3, the first six subgraphs (a)-(f) are the performance of the baseline algorithms. In the framework we designed, the parameters of the neural network are derived from the previous timestep, so

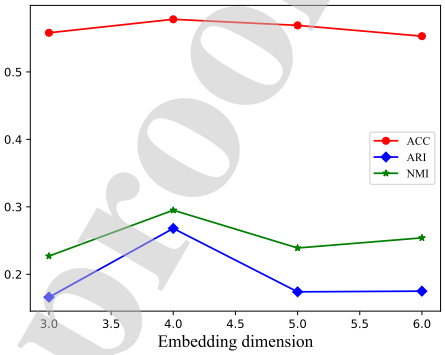


Figure 5: Recurrent-DC performance with different embedding dimensions in the Stacked Autoencoder.

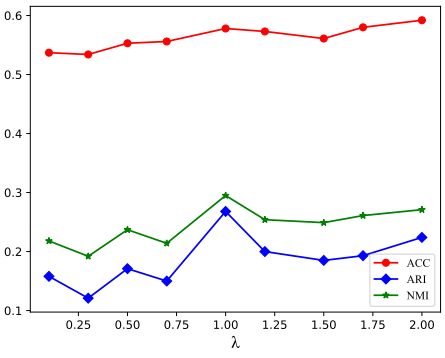


Figure 6: Recurrent-DC performance with different λ.

our effect can be continuously improved. Other comparison algorithms do not have a timestep and overall iterative process. The parameters are derived from one single learning process. For example, SAE+KM in the comparison algorithm, which performs k-means clustering after SAE training, so the effect will not continue to improve. It can be seen that the baseline algorithms do not take our recurrent process, so in the timestep of 10 to 100, there is no change in the evaluation metrics. By the end of timesteps, the Recurrent-DC algorithm we proposed achieved the best results in all three evaluation metrics. This result shows that the optimization algorithm work towards a desired direction. Recurrent-DC consistently achieves stable and significant performance over all timesteps.

4.2.2. Parameter sensitivity

Determining hyperparameters by cross-validation on a validation set is not an option in unsupervised clustering. In this subsection, we study the impacts of the following parameters: (1) the number of hidden layers of the Stacked Autoencoder, (2) learning rate, (3) embedding dimension, (4) parameter λ

Number of hidden layers of the Stacked Autoencoder. While the input data is determined, there are two important parameters to be determined accordingly: the number of input layer units and the number of output layer units. When we use the model, another question must be solved, that is how to optimize the number of hidden layers. Studies show the model will have a better performance with more hidden layers. However, the model will be prone to over-fitting, thus the cluster label prediction performance will be seriously reduced on the following clustering process. Since we need to consider the effective depth of the training so that the model can achieve better prediction without over-fitting.

In this paper, we need to determine hidden layers for the Stacked Autoencoder. For the Stacked Autoencoder, we choose the hidden layers from 1 to 5. The number of hidden units in each layer is presented in Table 2.

Table 2: The Number of hidden units in each layer for the Stacked Autoencoder

Stacked Autoencoder	Group	No. layers	No. units
Encoder	1	1	10
	2	2	10,30
	3	3	10,30,40
	4	4	10,20,30,40
	5	5	10,20,30,40,50
Decoder	1	1	10
	2	2	30,10
	3	3	40,30,10
	4	4	40,30,20,10
	5	5	50,40,30,20,10

Fig. 4 present the experimental results. From Fig. 4, we can observe that the group with three hidden layers obtained best results among three evaluation metrics. Thus we can obtain the best architecture of our proposed model with three hidden layers for the Stacked Autoencoder.

Learning rate. Learning rate is a crucial parameter in Recurrent-DC because it controls the update speed of the model. If we set the learning rate too large, the value of the loss function will move backwards and forwards around the minimum and will not converge. Otherwise it will lead to a slow learning process. In Table 3, we discover that if the learning rate is set to 0.001, the model can perform best on almost all metrics.

Table 3: Performance of Recurrent-DC with different learning rate

Learning rate	ACC	ARI	NMI
0.001	0.578	**0.268**	**0.295**
0.005	0.580	0.216	0.271
0.01	**0.595**	0.239	0.288
0.05	0.586	0.192	0.256
0.1	0.545	0.122	0.172

Embedding dimension. After learning, we will get the representation in terms of n dimensional vector, where n is the artificially set embedding dimension. To verify whether it has the impact on the performance of the model, we vary it from 3 to 6 in Recurrent-DC. The results are shown in Fig. 5. From the results we can see that, our model performs well with both lower and higher dimensional representation, which can be evidenced by the accuracy higher than baseline methods with different embedding dimensions.

Parameter λ selection. The parameter λ is important, since it trades off between the overall timestep and finding appropriate latent representations by random initialization. As we see from the experiments, the proposed Recurrent-DC works well with an appropriately chosen λ. Moreover, our experience suggests that the performance of our approach is insensitive to the exact value of λ. Fig. 6 shows how the proposed method performs with different λ on the dataset. As we can see, although there is a fluctuation of performance as λ gets large, the overall score trend becomes higher. The proposed method gives satisfactory result for a range of λ and nearly best result when λ close to one.

5. Application of University Profiling Framework

Taking advantage of the University Profiling Framework on the task of characterizing university, we can apply UPF on the MAG dataset to automatically obtain the university clusters. We select top universities according to ARWU as described in the Experiments section.

Table 4 shows the Research Complexity Index, Opportunity Value, Research Production Index and Productivity Value of

Table 4: RCI, OV, RPI, and PV of the selected universities

University Name	RCI	OV	RPI	PV	University Name	RCI	OV	RPI	PV
University of Cambridge	-0.75	-17.34	1.36	1.22	McMaster University	0.38	10.89	0.53	1.25
Univ of Southern California	-0.16	6.38	0.81	1.24	University of Edinburgh	-0.08	5.34	0.80	1.27
McGill University	0.51	15.40	0.87	1.14	University of Paris Sud	-0.87	-5.42	0.38	1.25
University of Munich	0.69	7.55	0.86	1.25	Boston University	0.70	11.01	0.63	1.31
Princeton University	-1.20	-54.77	0.55	1.52	Northwestern University	-0.19	3.93	1.07	1.30
Kyoto University	-0.85	-4.18	1.17	1.35	UNC	0.70	20.61	0.97	1.15
Karolinska Institute	2.87	8.08	0.59	1.06	Australian National University	-0.98	-27.30	0.45	1.42
University of Manchester	-0.41	-6.61	0.90	1.22	Leiden University	0.76	7.19	0.61	1.24
Yale University	0.50	13.95	1.10	1.27	California Institute of Technology	-1.51	-33.35	0.48	1.23
University of Florida	-0.07	7.85	0.95	1.10	Uppsala University	-0.06	4.04	0.57	1.17
ETH Zurich	-1.34	-43.97	0.66	1.26	University of Copenhagen	1.03	11.07	1.09	1.13
University of Bristol	-0.28	-0.23	0.60	1.40	University of Helsinki	0.94	9.75	0.62	1.07
Univ of California, San Diego	-0.37	0.08	1.11	1.29	University of Wisconsin - Madison	-0.67	-14.42	1.03	1.10
University of Paris VI	-0.85	-7.17	0.41	1.37	Harvard University	0.89	22.82	2.92	1.35
University of Oxford	-0.16	4.19	1.48	1.26	University of British Columbia	0.03	11.19	1.10	1.20
University of Chicago	-0.35	-1.83	0.78	1.30	Johns Hopkins University	2.10	14.22	1.76	1.25
University College London	0.61	17.16	1.37	1.28	Univ of California, Santa Barbara	-1.18	-50.43	0.33	1.30
Imperial College London	-0.31	1.05	1.08	1.32	UIUC	-1.06	-59.02	0.82	1.13
New York University	0.32	17.92	0.90	1.24	Univ of California, Los Angeles	0.47	15.97	1.34	1.24
Technical University Munich	-0.75	-10.09	0.74	1.20	Duke University	0.98	18.85	1.13	1.35
MIT	-1.15	-57.86	0.99	1.26	Univ of California, San Francisco	2.55	13.15	1.12	1.32
Univ of Maryland, College Park	-1.03	-54.54	0.63	1.28	King's College London	1.34	16.77	0.70	1.36
Rice University	-1.34	-39.06	0.28	1.08	University of Oslo	0.28	9.30	0.48	1.23
University of Tokyo	-0.97	-8.58	1.64	1.33	University of Toronto	0.49	20.42	1.55	1.24
Carnegie Mellon University	-1.18	-40.94	0.32	1.32	University of Colorado at Boulder	-1.10	-42.01	0.51	1.17
Columbia University	0.13	14.07	1.19	1.29	Utrecht University	-0.27	1.07	0.68	1.45
University of Pennsylvania	0.98	22.52	1.32	1.23	Vanderbilt University	1.18	10.05	0.62	1.21
Rockefeller University	1.27	2.32	0.10	1.17	University of Washington	0.56	20.02	1.44	1.19
University of California, Davis	-0.53	-4.98	0.85	1.12	University of Zurich	1.12	7.28	0.70	1.35
University of Texas at Austin	-1.01	-47.03	0.91	1.18	University of California, Irvine	-0.53	-6.45	0.58	1.23
Univ of California, Berkeley	-1.04	-68.64	1.08	1.15	Washington Univ in St. Louis	1.92	8.93	0.82	1.17
Stanford University	-0.84	-22.55	1.65	1.32	Cornell University	-0.86	-14.55	1.02	1.22

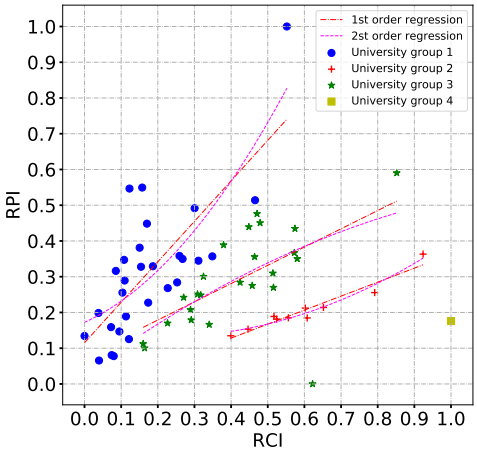


Figure 7: Relationship between normalized RPI and RCI of different university groups

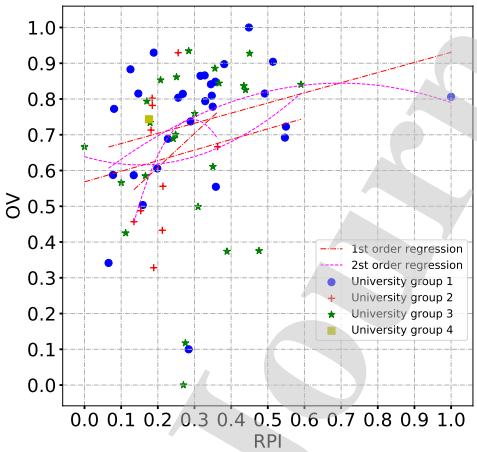


Figure 8: Scatter plot of normalized OV versus RPI of different university groups

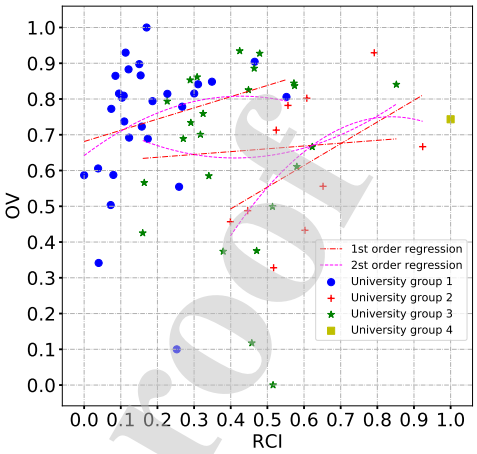


Figure 9: Scatter plot of normalized OV versus RCI of different university groups

selected universities. The order of the list follows the MAG 2018 university ID [1]. The top universities according to the research complexity RCI are Karolinska Institute (2.87), University of California, San Francisco (2.55), Johns Hopkins University (2.1), Washington University in St. Louis (1.92) and King's College London (1.34). The top universities according to the Opportunity Value are Harvard University (22.82), University of Pennsylvania (22.52), University of North Carolina at Chapel Hill (UNC) (20.61), University of Toronto (20.42), University of Washington (20.02). The top universities according to the research production RPI are Harvard University (2.92), Johns Hopkins University (1.76), Stanford University (1.65), University of Tokyo (1.64) and University of Toronto (1.55). The top universities according to the Productivity Value are Princeton University (1.52), Utrecht University (1.45), Australian National University (1.42), University of Bristol (1.40) and University of Paris VI (1.37).

The scatter plots of the Research Production Index versus the Research Complexity Index can be found in Fig. 7. As shown in the Fig. 7, these universities are divided into four clusters and each cluster represented by a different color. It should be noted that the yellow color cluster only contains Karolinska Institute. Similarly, the scatter plot of the Opportunity Value versus the Research Production Index is shown in Fig. 8, and the scatter plot of the Opportunity Value versus the Research Complexity Index is shown in Fig. 9. In these figures, different colors and shapes represent different cluster universities.

In Fig. 8 and Fig. 9, no obvious correlation between OV and RPI and RCI was observed. Different from Fig. 8 and Fig. 9, A high degree of correlation between Research Production Index and Research Complexity Index is observed in Fig. 7. As shown in the first and second-order regression, RPI well fit as a linear function of RCI of each cluster. And the second-order regression curve is not much different from the first-order regression line. This is maybe because universities with high

RCI values have more research fields within close proximity to purse or its research fields are considered complex so it is more likely to produce more papers. Thus, RCI could be considered a proxy of the Research Production Index which projects the potential production outlook in the future. The third-order regression is not included to increase the readability. It should be noted that different clusters of universities exhibit different rates of rising of correlation between Research Production Index and Research Complexity Index. This is maybe because different clusters of universities have different research patterns which could be due to a number of factors, such as research specialization, discipline-based research funds. These factors are beyond the scope of this paper and further investigations may be considered in the future work.

The research article classification may also affect the result of the Research Complexity Index. Different research articles classification standard may result in different university Revealed Symmetric Comparative Advantage. There are many different classification methods, like the SCImago Journal Classification, Web of Science subject category, Microsoft Academic Graph fields of study category. Comparing the Research Complexity Index under different research articles classification standards could be investigated. Addressing this problem is a feasible future work that is beyond the scope of this paper.

Further analysis, such as considering different university groups and different interpretable clustering methods, will be considered as the future work of this paper. Other intellectual property rights, such as patents, can be used as an alternative channel for disseminating university research. It is a common practice for academic institutions to attach research publications to patents. In this article [46], the national innovation capabilities based on patents are studied. Because patent applications have different motivations, this study does not consider patent records.

6. Conclusion

In this paper, we characterize top universities based on the quantification of complexity features in exploring academic graph. We propose an efficient University Profiling Framework for characterising scientific research institutions. Specially, we design a new deep representation clustering model, Recurrent-DC, to jointly learn representation and clustering. Experiments on the top university (according to ARWU) in the academic graph dataset demonstrate the effectiveness and efficiency of the proposed model. We propose new indicators including institutional complexity to describe university institutions. It is worth noting that the concept of institutional complexity that we have adopted is not complex in the traditional sense. The more disciplines of an institution and the stronger the competitiveness of disciplines, the higher the complexity of the institution. We divide the top universities into 4 categories based on the proposed indicators. Comparing the deep clustering algorithm we used with multiple comparison algorithms, our algorithm showed good results.

We conducted detailed experiments and discussions on the embedding dimension, learning rate and other parameters used by the deep clustering algorithm. Our work can provide new perspectives in the field of institutional profiling research and contribute to the academic evaluation of scientific research institutions. Often, institutions with high institutional research complexity have an advantage in scientific research competition. Moreover, we apply University Profiling Framework on the top university in academic graph dataset and explore the positive relationship between Research Production Index and Research Complexity Index. In the future work, we will investigate that how can we extend our deep representation clustering model to other types of data.

Acknowledgment

This work was partially supported by the National Natural Science Foundation of China under Grant No. 62072409 and 62073295.

References

[1] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B.-J. Hsu, K. Wang, An overview of microsoft academic service (mas) and applications, in: A. Gangemi, S. Leonardi, A. Panconesi (Eds.), Proceedings of the 24th International Conference on World Wide Web - WWW'15 Companion, ACM Press, New York, New York, USA, 2015, pp. 243–246.

[2] L. Bornmann, R. Mutz, H.-D. Daniel, Multilevel-statistical reformulation of citation-based university rankings: The leiden ranking 2011/2012, Journal of the American Society for Information Science and Technology 64 (8) (2013) 1649–1658.

[3] M. Dobrota, M. Bulajic, L. Bornmann, V. Jeremic, A new approach to the qs university ranking using the composite i-distance indicator: Uncertainty and sensitivity analyses, Journal of the Association for Information Science and Technology 67 (1) (2016) 200–211.

[4] X. Jiang, X. Sun, H. Zhuge, Towards an effective and unbiased ranking of scientific literature through mutual reinforcement, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM 12, Association for Computing Machinery, New York, NY, USA, 2012, pp. 714–723.

[5] J. A. García, R. Rodríguez-Sánchez, J. Fdez-Valdivia, N. Robinson-García, D. Torres-Salinas, Mapping academic institutions according to their journal publication profile: Spanish universities as a case study, Journal of the American Society for Information Science and Technology 63 (11) (2012) 2328–2340.

[6] I. Lee, Y. Tie, Fitness and research complexity among research-active universities in the world, IEEE Transactions on Emerging Topics in Computing (2018). doi:10.1109/TETC.2018.2854266.

[7] I. Lee, F. Xia, G. Roos, An observation of research complexity in top universities based on research publications, in: R. Barrett, R. Cummings, E. Agichtein, E. Gabrilovich (Eds.), Proceedings of the 26th International Conference on World Wide Web Companion - WWW'17 Companion, ACM Press, New York, New York, USA, 2017, pp. 1259–1265.

[8] S. Lloyd, Least squares quantization in pcm, IEEE Transactions on Information Theory 28 (2) (1982) 129–137.

[9] B. Yang, X. Fu, N. D. Sidiropoulos, M. Hong, Towards k-means-friendly spaces: Simultaneous deep learning and clustering, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 3861–3870.

[10] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion, Journal of Machine Learning Research 11 (Dec) (2010) 3371–3408.

[11] J. R. Hershey, Z. Chen, J. Le Roux, S. Watanabe, Deep clustering: Discriminative embeddings for segmentation and separation, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2016, pp. 31–35.

[12] J. Xie, R. Girshick, A. Farhadi, Unsupervised deep embedding for clustering analysis, in: Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48, 2016, pp. 478–487.

[13] F. Xia, W. Wang, T. M. Bekele, H. Liu, Big scholarly data: A survey, IEEE Transactions on Big Data 3 (1) (2017) 18–35.

[14] K. Williams, J. Wu, S. R. Choudhury, M. Khabsa, C. L. Giles, Scholarly big data information extraction and integration in the citeseer χ digital library, in: 2014 IEEE 30th International Conference on Data Engineering Workshops, IEEE, 2014, pp. 68–73.

[15] J. Priem, Beyond the paper, Nature 495 (7442) (2013) 437–440.

[16] X. Kong, Y. Shi, S. Yu, J. Liu, F. Xia, Academic social networks: Modeling, analysis, mining and applications, Journal of Network and Computer Applications 132 (2019) 86–103.

[17] T. Pradhan, S. Pal, A hybrid personalized scholarly venue recommender system integrating social network analysis and contextual similarity, Future Generation Computer Systems (2019). doi:10.1016/j.future.2019.11.017.

[18] S. Aslan, M. Kaya, Topic recommendation for authors as a link prediction problem, Future Generation Computer Systems 89 (2018) 249–264.

[19] Z. Wu, J. Wu, M. Khabsa, K. Williams, H.-H. Chen, W. Huang, S. Tuarob, S. R. Choudhury, A. Ororbia, P. Mitra, et al., Towards building a scholarly big data platform: Challenges, lessons and opportunities, in: Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries, IEEE Press, 2014, pp. 117–126.

[20] L. Bornmann, F. de Moya Anegón, R. Mutz, Do universities or research institutions with a specific subject profile have an advantage or a disadvantage in institutional rankings?, Journal of the American Society for Information Science and Technology 64 (11) (2013) 2310–2316.

[21] E. A. Corrêa Jr, F. N. Silva, L. d. F. Costa, D. R. Amancio, Patterns of authors contribution in scientific manuscripts, Journal of Informetrics 11 (2) (2017) 498–510.

[22] D. R. Amancio, O. N. Oliveira Jr, L. d. F. Costa, On the use of topological features and hierarchical characterization for disambiguating names in collaborative networks, EPL (Europhysics Letters) 99 (4) (2012) 48002.

[23] S. Pan, R. Hu, S.-f. Fung, G. Long, J. Jiang, C. Zhang, Learning graph embedding with adversarial training methods, IEEE Transactions on Cybernetics 50 (6) (2019) 2475–2487.

[24] S. Ji, S. Pan, E. Cambria, P. Marttinen, P. S. Yu, A survey on knowledge graphs: Representation, acquisition and applications, arXiv preprint arXiv:2002.00388 (2020).

[25] S. Wan, C. Gong, P. Zhong, S. Pan, G. Li, J. Yang, Hyperspectral image classification with context-aware dynamic graph convolutional network, IEEE Transactions on Geoscience and Remote Sensing (2020). doi:10.1109/TGRS.2020.2994205.

[26] T. Guo, S. Pan, X. Zhu, C. Zhang, Cfond: consensus factorization for co-clustering networked data, IEEE Transactions on Knowledge and Data Engineering 31 (4) (2018) 706–719.

[27] X. Cai, J. Han, W. Li, R. Zhang, S. Pan, L. Yang, A three-layered mutually reinforced model for personalized citation recommendation, IEEE transactions on neural networks and learning systems 29 (12) (2018) 6026–6037.

[28] J. MacQueen, et al., Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Oakland, CA, USA, 1967, pp. 281–297.

[29] B. J. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (5814) (2007) 972–976.

[30] A. Y. Ng, M. I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: Advances in Neural Information Processing Systems, 2002, pp. 849–856.

[31] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al., A density-based algorithm for discovering clusters in large spatial databases with noise., in: KDD, 1996, pp. 226–231.

[32] S. Guha, R. Rastogi, K. Shim, Cure: an efficient clustering algorithm for large databases, in: ACM Sigmod Record, ACM, 1998, pp. 73–84.

[33] T. Zhang, R. Ramakrishnan, M. Livny, Birch: an efficient data clustering method for very large databases, in: ACM Sigmod Record, ACM, 1996, pp. 103–114.

[34] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.

[35] V. Le Quoc, A. Ranzato Marc, R. Monga, M. Devin, K. Chen, S. Corrado Greg, et al., Building high-level features using large scale unsupervised learning, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 8595–8598.

[36] F. Tian, B. Gao, Q. Cui, E. Chen, T.-Y. Liu, Learning deep representations for graph clustering, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI14, AAAI Press, 2014, pp. 1293–1299.

[37] J. Yang, D. Parikh, D. Batra, Joint unsupervised learning of deep representations and image clusters, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5147–5156.

[38] J. L. Ortega, I. F. Aguillo, Microsoft academic search and google scholar citations: Comparative analysis of author profiles, Journal of the Association for Information Science and Technology 65 (6) (2014) 1149–1156.

[39] X. Kong, J. Zhang, D. Zhang, Y. Bu, Y. Ding, F. Xia, The gene of scientific success, ACM Transactions on Knowledge Discovery from Data (TKDD) 14 (4) (2020).

[40] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Networks 2 (5) (1989) 359–366.

[41] D. Cai, X. He, J. Han, Locally consistent concept factorization for document clustering, IEEE Transactions on Knowledge and Data Engineering 23 (6) (2010) 902–913.

[42] V. Nair, G. E. Hinton, Rectified linear units improve restricted boltzmann machines, in: Proceedings of the 27th International Conference on International Conference on Machine Learning, 2010, pp. 807–814.

[43] M. Thelwall, Microsoft academic automatic document searches: Accuracy for journal articles and suitability for citation analysis, Journal of Informetrics 12 (1) (2018) 1–9.

[44] J. Liu, F. Xia, L. Wang, B. Xu, X. Kong, H. Tong, I. King, Shifu2: A network representation learning based model for advisor-advisee relationship mining, IEEE Transactions on Knowledge and Data Engineering (2019). doi:10.1109/TKDE.2019.2946825.

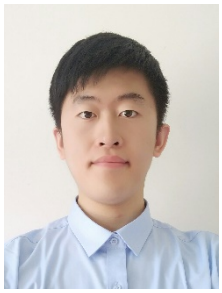
[45] M. Thelwall, Microsoft academic: A multidisciplinary comparison of citation counts with scopus and mendeley for 29 journals, Journal of Informetrics 11 (4) (2017) 1201–1212.

[46] J. L. Furman, M. E. Porter, S. Stern, The determinants of national innovative capacity, Research Policy 31 (6) (2002) 899–933.

Highlights

- We propose a deep representation clustering model based on academic graph.
- We propose a university profiling framework that transforms the traditional depiction of excellence into the quantification of complexity indicators.
- We find the positive relationship between university research production and university research complexity of different university groups.
- Experimental results show that the model achieves the state-of-the-art performance.

Xiangjie Kong



Jiaxing Li

Luna Wang

Guojiang Shen

Yiming Sun

Ivan Lee

Xiangjie Kong received the B.Sc. and Ph.D. degrees from Zhejiang University, Hangzhou, China. He is currently a Full Professor with College of Computer Science and Technology, Zhejiang University of Technology. Previously. Previously, he was an Associate Professor with the School of Software, Dalian University of Technology, China. He has published over 130 scientific papers in international journals and conferences (with over 100 indexed by ISI SCIE). His research interests include network science, mobile computing, and computational social science. He is a Senior Member of the IEEE and CCF and is a member of ACM.

Jiaxing Li received the B.Sc. degree from Northwest A&F University, China, in 2018. He is currently working toward the master's degree in the School of Software, Dalian University of Technology, China. His research interests include deep learning, social computing, and data science.

Luna Wang received M.Sc. degree from Zhejiang University, China. She is currently working in Institute of Science and Technology, Dalian University of Technology. Her research interests include educational big data and knowledge management.

Guojiang Shen received the B.Sc. degree in control theory and control engineering and the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 1999 and 2004, respectively. He is currently a Professor with College of Computer Science and Technology, Zhejiang University of Technology. His current research interests include artificial intelligence theory, big data analytics, and intelligent transportation systems.

Yiming Sun is currently an undergraduate of Software Engineering, Dalian University of Technology, China. She is currently working towards the B.Sc. degree. Her research interests include big scholarly data and network science.

Ivan Lee received BEng, MCom, MER, and PhD degrees from the University of Sydney, Australia. He was a software development engineer at Cisco Systems, a software engineer at Remotek Corporation, and an assistant professor at Ryerson University. Since 2008, he has been a senior lecturer at the University of South Australia. His research interests include smart sensors, multimedia systems, and data analytics.

Author Statement

Xiangjie Kong: Conceptualization, Supervision, Writing - Original Draft. **Jiaxing Li**: Methodology, Software, Writing - Review & Editing. **Luna Wang**: Investigation, Writing - Review & Editing. **Guojiang Shen**: Validation, Writing - Review & Editing. **Yiming Sun**: Software, Writing - Review & Editing. **Ivan Lee**: Supervision, Writing - Review & Editing.

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: